PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 :
G06F 17/30

A1

(11) International Publication Number: WO 00/08573

(43) International Publication Date: 17 February 2000 (17.02.00)

(21) International Application Number: PCT/US99/17654

(22) International Filing Date: 4 August 1999 (04.08.99)

(30) Priority Data:
60/095,296 4 August 1998 (04.08.98) US

(71) Applicant: RULESPACE, INC. [US/US]; Suite 400, 208 Southwest Stark Street, Portland, OR 97204 (US).

(72) Inventor: KAWASAKI, Charles; 8046 Northwest Blue Pointe Lane, Portland, OR 97229 (US).

(74) Agent: STOLOWITZ, Micah, D.; Stoel Rives LLP, Suite 2600, 900 S.W. Fifth Avenue, Portland, OR 97204-1268 (US).

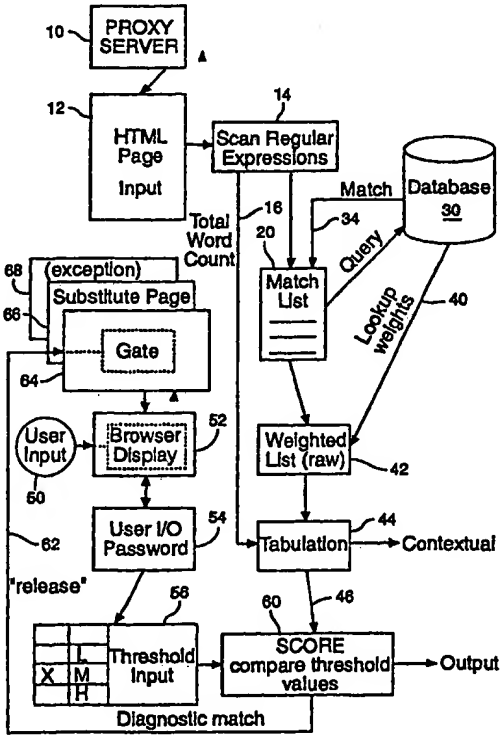
(81) Designated States: AE, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CR, CU, CZ, DE, DK, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, UZ, VN, YU, ZA, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published
With international search report.

(54) Title: METHOD AND SYSTEM FOR DERIVING COMPUTER USERS' PERSONAL INTERESTS

(57) Abstract

A method and system of this invention that provides for profiling a user of the Internet according to predefined categories of interest is described herein (Figure 5). This invention includes scanning content information (Figures 5, 14) of an Internet user to generate unknown data (42) and processing unknown data to determine its relevance to predefined categories of interest (Figures 5, 60). This relevance is indicated in a matching rating system (Figures 5, 56), analyzing attributes such as length of time reviewing content information, frequency of encounter, recency, strength and closeness (Figures 6, 7 and 8).



FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece			TR	Turkey
BG	Bulgaria	HU	Hungary	ML	Mali	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MN	Mongolia	UA	Ukraine
BR	Brazil	IL	Israel	MR	Mauritania	UG	Uganda
BY	Belarus	IS	Iceland	MW	Malawi	US	United States of America
CA	Canada	IT	Italy	MX	Mexico	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NE	Niger	VN	Viet Nam
CG	Congo	KE	Kenya	NL	Netherlands	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NO	Norway	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	NZ	New Zealand		
CM	Cameroon			PL	Poland		
CN	China	KR	Republic of Korea	PT	Portugal		
CU	Cuba	KZ	Kazakstan	RO	Romania		
CZ	Czech Republic	LC	Saint Lucia	RU	Russian Federation		
DE	Germany	LI	Liechtenstein	SD	Sudan		
DK	Denmark	LK	Sri Lanka	SE	Sweden		
EE	Estonia	LR	Liberia	SG	Singapore		

METHOD AND SYSTEM FOR DERIVING COMPUTER USERS' PERSONAL INTERESTS

Field of the Invention

This invention relates to the Internet and more particularly to a method and system for monitoring the use of the Internet by users and generating profile data for use in targeting users according to their interests.

Background of the Invention

Users of Internet services ("Users") now include families, children, business people, students, hobbyists and enthusiasts of all types. They use the Internet from home, work and school.

Users of Internet services ("Services") are rapidly becoming familiar with (and are beginning to expect) new Services for "free." These "free" systems and Services are able to provide low-cost applications and services by supporting their infrastructures through the use of, and exploitation of, large audiences. The Services provide an infrastructure into which marketing and advertising companies ("Advertisers") market their products and services through the placement of on-line promotional offers and on-line advertisements ("Offers"). This model is similar to the ubiquitous network TV model of free programming, with commercial breaks. Even paid-for Services on the Internet have begun "mining" this source of value, through placing advertisements in key locations. Offers have begun to take on a variety of forms, including Web-based banner ads, e-mails, pop-up screens and video

interstitials. Additionally, Services have begun to use more traditional means to target Offers, including print campaigns, radio and direct mail.

Examples of these Services on the Internet include "free" search engines and directories such as Yahoo and Infoseek, "free" e-mail systems such as HotMail, "free" instant messaging systems such as ICQ and "free" information broadcast systems such as PointCast Network.

The revenue generated by these Services in 1997 was nearly $1 billion and was in large part generated by charging Advertisers on a "thousands of impressions" model. This is commonly understood in the advertising industry as selling advertisements by "CPM" - a measure of 1,000 "impressions." Impressions are counted as one impression for each time a potential customer of the Advertiser's product sees the Offer. This is commonly understood in the Internet industry as "page views." Common "page view" prices currently range from $10 to $400 per 1,000 CPM (one million page views).

The CPM price varies widely, based on the appropriateness (or demographic match) of the User to the Offer. For example, the CPM price for an Internet search service that caters to the broadest categories of Users may command a very low CPM such as $20. The CPM price for a highly targeted audience, such as a membership-based medical information Web site that has a well-known group of subscribers suffering from a specific disease, may run as high as $450. The economics of these models work to drive specific Advertisers to specific Services. For example, drug and health product Advertisers can justify paying $450 per CPM on health-specific sites, because they are able to reach highly targeted audiences that have a great match and affinity to their Offer.

However, the economic model presented by many Services to Advertisers is highly inefficient. Advertising on Services such as Web sites generally generate between a 1% and 10% response rate, with 2% as a typical response rate. There are numerous reasons for poor response rates such as 2%, including poor or unappealing designs of the Offer, but one of the main reasons for poor response rates is simply that an Offer is tendered to Users who have no interest in the product, service and/or

subject matter. For example, Advertisers of golfing equipment or luxury automobiles have little assurance that placing Offers on Services such as CNN.com or Time Warner's pathfinder.com will result in their messages reaching a high percentage of their target audiences.

Services can provide extremely attractive mediums for Advertisers by improving the match between the subject of the Offer and a User's interests. A small improvement in response rate for Advertisers, even as small as 2%, can substantially improve the economics of placing Offers in Services. Thus Advertisers and Services have great interest in techniques of measuring and improving the match of Offer to User. These techniques are known as "Offer targeting." With improved Offer targeting, Services can profit dramatically by sharing in the improved performance by increasing the CPM price for their systems.

Some Services have attempted to improve the efficiency of their Offer targeting through manual means. These have included manual organization of Services into known topic areas that Advertisers may select to place offers into. This strategy works somewhat when the content delivered by Services is well known and easily classed into categories or is static in nature. This approach does not work well for a dynamic medium with huge sets of rapidly changing content and the content is out of the control of the Service – which are attributes of the World Wide Web.

Other targeting techniques include requiring Users to specify their interest categories manually. These systems may work for small numbers of well-understood information content areas but are not practical for Services that span the breadth of the Internet. Furthermore, manual systems of specifying preferences are cumbersome for Users, which they subsequently abandon. This results in inaccurate preferences, misleading preferences or obsolete preferences, causing a mismatch between actual User interests and the information captured in manual preferences systems.

Summary of the Invention

In view of the foregoing background, one object of the present invention is to improve the match between User interests and Advertisers' messages by transparently assessing the type of information that a User reviews over time.

Another object of the invention is to create a "profile" of interests of the User, which may be used to subsequently direct Offers.

A further object of the invention is to target Offers to individuals who have indirectly expressed interest in specific subject matter.

A more general object of this invention is to provide a higher User response rate per CPM.

According to one aspect of the invention, a method and system of this invention provides for profiling a User of the Internet according to predefined categories of interest that includes the following steps: First, content information of an Internet User is scanned to generate unknown data. This step takes place at any number of locations: the client's server, the client's computer or at an Internet hub. Next, the unknown data are processed to determine their relevance to predefined categories of interest. These categories include, for example, sports, games, business, investing, health, hobbies, technology, arts, politics, social issues, weather and news. Moreover, the relevance is indicated in a matching rating system, analyzing attributes such as length of time reviewing content information, frequency of encounter, recency, strength and closeness. With comparisons such as these, the method generates a match of the unknown data with the predefined categories to form a profile of the User.

To form a "recognizer" for use in profiling Internet User interests, the method and system of this invention includes collecting representative data sets of major areas of interests and processing the data sets by algorithms and weighted rules to form a recognizer. The above-described profiling may occur in real time or be delayed and may occur on the client's installation or remotely, for example, on a server installation.

Therefore, the objects of this invention are accomplished through a method system that scans information content and automatically and transparently assesses its subject matter. Over time, this invention accumulates a "profile" of interests of the User, which can be used to subsequently direct Offers.

An advantage of this invention is that it enables Services to target offers to individuals who have indirectly expressed interest in specific subject matter. For example, this will enable Services to automatically and transparently determine which Users have an interest in topics such as golf, luxury cars, medical information, sports equipment, music, etc. These topics are merely illustrative and not limiting. Once the Service has determined these preferences, it may then direct appropriate and matching advertising Offers to those Users.

Another advantage of this invention will be a higher response rate per CPM, i.e., golf advertisements will be shown only to those Service Users who have a strong interest in golf, as determined in the profile generated by Petitioner's technology.

Additional objects and advantages of this invention will be apparent from the following detailed description of preferred embodiments thereof, which proceeds with reference to the accompanying drawings.

Brief Description of the Drawings

Fig. 1 shows the relationship between the Recognizers, the User and the Profiler according to this invention in connection with e-mail, Web and Push data streams.

Fig. 2 shows the relationship between the Data Sets, Neural Net Processing and the Recognizers of this invention.

Fig. 3 is a flow chart of the operation to form a User profile of this invention.

Fig. 4 shows the relationship between the Recognizers, the User, the Profiler and other components of the system and method of this invention.

Fig. 5 is a flow diagram illustrating operation of a process for blocking display of a web page or other digital dataset that contains a particular type of content such as pornography.

Fig. 6 is a simplified block diagram of a modified neural network architecture for creating a weighted list of regular expressions useful in analyzing content of a digital dataset.

Fig. 7 is a simplified diagram illustrating a process for forming a target attribute set having terms that are indicative of a particular type of content, based on a group of training datasets.

Fig. 8 is a flow diagram illustrating a neural network based adaptive training process for developing a weighted list of terms useful for analyzing content of web pages or other digital datasets.

Detailed Description of a Preferred Embodiment of the Invention

This invention dynamically and transparently improves the targeting of Offers consisting of three major components. These three components work in conjunction with technologies of Services to target Offers to Users. Two of these components, scanning and handling various kinds of digital information content, are described below beginning at page 12 and in FIGS. 5-8. Referring to FIG. 1, the scanning and analyzing capabilities extend to any type of digital content in systems such as the Web Browsers 11, E-mail Clients/Servers 12, UseNet Clients/Servers 13, Personal 14 and Server-based Search Engines 16.

The two modules, scanning and analyzing, named in the above-referenced patent application are capable of determining the relevance of unknown data to a known data set through efficient analytic models. These modules are used in real time to assess the incoming data from Services to the Users of the Services for their relevance to the common (or specific) predefined categories of interest to Advertisers. Referring to FIG. 2, this is accomplished through collecting representative data sets of major areas of interest, the Data Sets 21, and using a developed set of algorithms and weighted rules necessary to analyze the unknown content for a match with Data Sets 21. The development of these algorithms and weighted rules is accomplished through the use of a three-tier feed-forward artificial neural network, a Neural Net 22, with a learning algorithm as described in the above-referenced patent

application. Various artificial neural networks are commercially available that could be used for this purpose. The output from Neural Net 22 is algorithms and rules, the Recognizers 23, which in essence "recognize" a match of incoming, unknown data with Data Sets 21.

Recognizers 23 for common areas of interest to Advertisers include, but are not limited to, data sets related to sports, games, business, investing, health, hobbies, technology, arts, politics, social issues, weather and news. In addition, because Recognizers 23 are very small and compact, dynamically generated Recognizers 23 are used for electronic distribution and updates.

Recognizers are executed against incoming unknown data requested by Users. Again referring to FIG. 1, the statistical output from Recognizers 23 indicate whether a given set of unknown data received or sent through the Internet 18 has a good match to the installed Recognizers 23. For example, if a golfing Recognizer 23 is loaded and the User views golf-related Web Pages 13a, E-mail 12a, User Groups 16a or other digital content, the golfing Recognizer 23 returns a positive match for that Data Set 21.

Another element of this invention is the Profiler 26 that receives the output from the real-time analysis of Recognizers 23. Profiler 26 tracks and builds an aggregate statistical model, "Profile," of the User including quantitative analysis of the match, frequency, duration, age and other factors between unknown content reviewed by the User and the set of installed Recognizers 23.

The aggregate result of generating Profiler 26 is a prioritized and rated set of interest categories that is automatically generated for each User through the transparent and dynamic analysis of the frequency and time spent by the Users reviewing content that has a good statistical match with known Data Sets 21. The Profile generated by the system for each User is the output from the system and can subsequently be used to make highly targeted Offers.

Referring to FIG. 3, the Offer-managing software can use an aggregate of the interest ratings generated by Profiler 26 for each User, along with statistics generated

by the Offer-managing software, to report the level of targeted Offers back to the Advertiser.

The method and system of this invention employs methods for scanning, analyzing and handling digital information content, described below beginning at page 12 and in FIGS. 5-8 to assess the match between known Data Sets 21 of subject matter and newly encountered information.

Referring to FIG. 3, each discrete unit of information newly encountered by Users is analyzed by Recognizers 23 against a set of known Data Sets 21 such as sports, news, heath information, etc. The return value of the analysis is a matching rating indicating the "strength" or "closeness" of the newly encountered data to known Data Sets 21. This analysis is done for each known Data Set 21 (or subject matter) of interest to the Service and Advertiser.

Additional information is captured regarding the use of such data by the User. Information including a frequency of encounter 31, a length of time reviewing 32 and a statistical measure of how recently the matching information was reviewed, along with an aging algorithm 33, among other criteria 34, is used to generate a "level of recent interest" rating 36 for each known Data Set 21. Optionally, a history may be recorded, including the location of the newly encountered information, for use in subsequent validation of results.

The aggregate of this collected data, on a per data set basis, is mathematically combined to a single rating of "interest" level for each Data Set 21 for each User. These ratings can be sorted by highest "interest" first, through data sets of "No" interest at step 37. Referring to FIG. 4, this sorting can be done in real time to generate a profile 38 and reported out of the "Tracker" module incorporated with or in communication with the offer manager 41 into other modules responsible for delivering the Offers 42 that match the subject matter of highest interests.

Still referring to FIG. 4, the Services 13, 16 incorporating Profiler 38 and 41 identify which User is using the system. Profiles are preferably generated on a per-User basis. Typical installations require Users to "log in" to the Service, thereby allowing the Service to notify the "Tracker" which User Profile to update.

Services 13, 16 pass the discrete pieces of data into Recognizer 23 for the system to generate the appropriate rating data. Services 13, 16 may "tap" into the data at many different sources, including both realtime and delayed. Services 13, 16 may intercept communications traffic at the protocol or file layer of client and server platforms.

Profiles generated by Profiler 38 describe a graph of interest for each User. These are stored either on the client installation or the server installation and may or not be encrypted, depending upon the desires and privacy policies set by Services 13, 16 using the information. Additionally, personal information such as a user name, address, phone number, and other various forms of personal identification may be stripped from the profile to protect the user's anonymity. Full copies of the graphs are available for use by Services 13, 16, but privacy policies may dictate that only aggregate Profile rating is made available.

Users may have some options to review the history and aggregate results and may have the option to explicitly turn off tracking of specific subject matter. For example, Users may wish to disallow any reporting on frequency of use and may also flush the histories and graphs to reset the profiling functions.

The following is a sample list of applications for which the system and method of this invention are used:

1. E-mail Client/Server Systems - Analyzing and building Profiles from e-mail information. According to this invention, e-mail systems generate revenue through Offers by analyzing sent and received e-mail information and building Profiles based upon its content.

Analysis and Profile building can be accomplished at either the Client or Server location within an e-mail system.

E-mail systems include LANs, WANs, VPNs and ISPs that deploy e-mail information in any form of e-mail data standard such as SMTP, POP3, IMAP4, etc.

Clients refers to any software loaded on desktop PCS, set top boxes and end-user display devices that display e-mail.

Proxies and servers refer to any centralized computer system responsible for serving, routing, filtering and/or caching e-mail content.

2. Web Client/Proxy and Server - Analyzing and building Profiles from Web pages. This invention operates in conjunction with Web-based systems to generate revenue through Offers by analyzing viewed Web pages and building Profiles based upon their content.

Web systems include LANs, WANs, VPNs and ISPs that deploy Web information in the form of HTML, XML and other Web data standards.

Clients refers to any software loaded on desktop PCS, set top boxes and end-user display devices that display HTML, XML and other Web data.

Proxies and servers refer to any centralized computer system responsible for serving, routing, filtering and/or caching Web content.

3. Chat Client/Proxy and Server - Analyzing and building Profiles from chat streams. This invention can also be used in conjunction with chat-based systems to generate revenue through Offers by analyzing chat discussions and building Profiles based upon their content.

Chat systems include LANs, WANs, VPNs and ISPs that deploy chat information in the form of HTML, Java, TCP/IP, IRC or UDP chat protocols.

Clients refers to any software loaded on desktop PCS, set top boxes and end-user display devices that display chat conversations.

Proxies and servers refer to any centralized computer system responsible for serving, routing, filtering and/or caching chat content.

4. UseNet Client/Proxy and Server - Analyzing and building Profiles from UseNet information. Another example of an application of this invention is in conjunction with UseNet-based systems to generate revenue through Offers by analyzing viewed UseNet postings and building Profiles based upon their content.

UseNet systems include LANs, WANs, VPNs and ISPs that deploy UseNet information in the form of NNTP.

Clients refers to any software loaded on desktop PCS, set top boxes and end-user display devices that display UseNet data.

Proxies and servers refer to any centralized computer system responsible for serving, routing, filtering and/or caching UseNet content.

5. "Push" Client and Server - Analyzing and building Profiles from "Push" information. Push-based systems can be used in conjunction with this invention to generate revenue through Offers by analyzing viewed Push information and building Profiles based upon its content.

Push systems include LANs, WANs, VPNs and ISPs that deploy Push information in the form of channels, typically utilizing HTML or TCP/IP protocols.

Clients refers to any software loaded on desktop PCS, set top boxes and end-user display devices that display HTML, XML, "ticker" and other Push data.

Proxies and servers refer to any centralized computer system responsible for serving, routing, filtering and/or caching Push content.

6. "Portal" Client and Server - Analyzing and building Profiles from Web page information. Also, Web-based "Portal" or "Community" systems can be used in conjunction with this invention to generate revenue through Offers by analyzing viewed Web pages and building Profiles based upon their content.

Portal systems include LANs, WANs, VPNs and ISPs that deploy Web information in the form of HTML, XMI and other Web data standards, for the purpose of organizing and directing the Web experience for the User.

Clients refers to any software loaded on desktop PCS, set top boxes and end-user display devices that display HTML, XML and other Web data.

Proxies and servers refer to any centralized computer system responsible for serving, routing, filtering and/or caching Portal content.

7. Search Engine Client and Server - Analyzing and building Profiles from Web page information. Additionally, this invention can be used in conjunction with search engine systems to generate revenue through Offers by analyzing viewed search queries and results and building Profiles based upon their content.

Search engine systems include LANs, WANs, VPNs and ISPs that deploy search queries and results in the form of HTML, XML and other Web data standards.

Clients refers to any software loaded on desktop PCS, set top boxes and end-user display devices that display HTML, XML and other Web data related to search queries and results.

Proxies and servers refer to any centralized computer system responsible for serving, routing, filtering and/or caching search queries and results.

It will be apparent to those having skill in the art that many changes may be made to the details of the above-described embodiment of this invention without departing from the underlying principles thereof. The scope of the present invention should, therefore, be determined by the claims.

METHOD FOR SCANNING, ANALYZING AND HANDLING VARIOUS KINDS OF DIGITAL INFORMATION CONTENT

The technology is useful for a variety of applications, including but not limited to blocking digital content, especially world-wide web pages, from being displayed when the content is unsuitable or potentially harmful to the user, or for any other reason that one might want to identify particular web pages based on their content.

According to one aspect of the technology, a method for controlling access to potentially offensive or harmful web pages includes the following steps: First, in conjunction with a web browser client program executing on a digital computer, examining a downloaded web page before the web page is displayed to the user. This examining step includes identifying and analyzing the web page natural language content relative to a predetermined database of words -- or more broadly regular expressions -- to form a rating. The database or "weighting list" includes a list of expressions previously associated with potentially offensive or harmful web pages, for example pornographic pages, and the database includes a relative weighting assigned to each word in the list for use in forming the rating.

The next step is comparing the rating of the downloaded web page to a predetermined threshold rating. The threshold rating can be by default, or can be selected, for example based on the age or maturity of the user, or other

"categorization" of the user, as indicated by a parent or other administrator. If the rating indicates that the downloaded web page is more likely to be offensive or harmful than a web page having the threshold rating, the method calls for blocking the downloaded web page from being displayed to the user. In a presently preferred embodiment, if the downloaded web page is blocked, the method further calls for displaying an alternative web page to the user. The alternative web page can be generated or selected responsive to a predetermined categorization of the user like the threshold rating. The alternative web page displayed preferably includes an indication of the reason that the downloaded web page was blocked, and it can also include one or more links to other web pages selected as age-appropriate in view of the categorization of the user. User login and password procedures are used to establish the appropriate protection settings.

Of course the technology is fully applicable to digital records or datasets other than web pages, for example files, directories and email messages. Screening pornographic web pages is described to illustrate the technology and it reflects a commercially available embodiment of the technology.

Another aspect of the technology is a computer program. It includes first means for identifying natural language textual portions of a web page and forming a list of words or other regular expressions that appear in the web page; a database of predetermined words that are associated with the selected characteristic; second means for querying the database to determine which of the list of words has a match in the database; third means for acquiring a corresponding weight from the database for each such word having a match in the database so as to form a weighted set of terms; and fourth means for calculating a rating for the web page responsive to the weighted set of terms, the calculating means including means for determining and taking into account a total number of natural language words that appear in the identified natural language textual portions of the web page.

As alluded to above, statistical analysis of a web page according to the technology requires a database or attribute set, compiled from words that appear in know "bad" -- e.g. pornographic, hate-mongering, racist, terrorist, etc. -- web pages.

The appearance of such words in a downloaded page under examination does not necessarily indicate that the page is "bad", but it increases the probability that such is the case. The statistical analysis requires a"weighting" be provided for each word or phrase in a word list. The weightings are relative to some neutral value so the absolute values are unimportant. Preferably, positive weightings are assigned to words or phrases that are more likely to (or even uniquely) appear in the selected type of page such as a pornographic page, while negative weightings are assigned to words or phrases that appear in non-pornographic pages. Thus, when the weightings are summed in calculating a rating of a page, the higher the value the more likely the page meets the selected criterion. If the rating exceeds a selected threshold, the page can be blocked.

A further aspect of the technology is directed to building a database or target attribute set. Briefly, a set of "training datasets" such as web pages are analyzed to form a list of regular expressions. Pages selected as "good" (non-pornographic, for example) and pages selected as "bad" (pornographic) are analyzed, and rate of occurrence data is statistically analyzed to identify the expressions (e.g. natural language words or phrases) that are helpful in discriminating the content to be recognized. These expressions form the target attribute set.

Then, a neural network approach is used to assign weightings to each of the listed expressions. This process uses the experience of thousands of examples, like web pages, which are manually designated simply as "yes" or "no" as further explained later.

Additional objects and advantages of this technology will be apparent from the following detailed description of preferred embodiments thereof which proceeds with reference to the accompanying drawings.

Brief Description of the Drawings

Fig. 5 is a flow diagram illustrating operation of a process according to the present technology for blocking display of a web page or other digital dataset that contains a particular type of content such as pornography.

Fig. 6 is a simplified block diagram of a modified neural network architecture for creating a weighted list of regular expressions useful in analyzing content of a digital dataset.

Fig. 7 is a simplified diagram illustrating a process for forming a target attribute set having terms that are indicative of a particular type of content, based on a group of training datasets.

FIG. 8 is a flow diagram illustrating a neural network based adaptive training process for developing a weighted list of terms useful for analyzing content of web pages or other digital datasets.

Detailed Description of a Preferred Embodiment

FIG. 5 is a flow diagram illustrating operation of a process for blocking display of a web page (or other digital record) that contains a particular type of content. As will become apparent from the following description, the methods and techniques of the present technology can be applied for analyzing web pages to detect any specific type of selected content. For example, the technology could be applied to detect content about a particular religion or a particular book; it can be used to detect web pages that contain neo-Nazi propaganda; it can be used to detect web pages that contain racist content, etc. The presently preferred embodiment and the commercial embodiment of the technology are directed to detecting pornographic content of web pages. The following discussions will focus on analyzing and detecting pornographic content for the purpose of illustrating the technology.

In one embodiment, the technology is incorporated into a computer program for use in conjunction with a web browser client program for the purpose of rating web pages relative to a selected characteristic—pornographic content, for example—and potentially blocking display of that web page on the user's computer if the content is determined pornographic. In FIG. 5, the software includes a proxy server 10 that works upstream of and in cooperation with the web browser software to receive a web page and analyze it before it is displayed on the user's display

screen. The proxy server thus provides an HTML page 12 as input for analysis. The first analysis step 14 calls for scanning the page to identify the regular expressions, such as natural language textual portions of the page. For each expression, the software queries a pre-existing database 30 to determine whether or not the expression appears in the database. The database 30, further described later, comprises expressions that are useful in discriminating a specific category of information such as pornography. This query is illustrated in FIG. 5 by flow path 32, and the result, indicating a match or no match, is shown at path 34. The result is formation of a "match list" 20 containing all expressions in the page 12 that also appear in the database 30. For each expression in the match list, the software reads a corresponding weight from the database 30, step 40, and uses this information, together with the match list 20, to form a weighted list of expressions 42. This weighted list of terms is tabulated in step 44 to determine a score or rating in accordance with the following formula:

$$\text{rating} = (n\sum_{1}^{p}(x_p w_p)) / c$$

In the above formula, "n" is a modifier or scale factor which can be provided based on user history. Each term x_p w_p is one of the terms from the weighted list 42. As shown in the formula, these terms are summed together in the tabulation step 44, and the resulting sum is divided by a total word count provided via path 16 from the initial page scanning step 14. The total score or rating is provided as an output at 46.

Turning now to operation of the program from the end-user's perspective, again referring to FIG. 5, the user interacts with a conventional web browser program by providing user input 50. Examples of well-known web-browser programs include Microsoft Internet Explorer and Netscape. The browser displays information through the browser display or window 52, such as a conventional PC monitor screen. When the user launches the browser program, the user logs-in for present purposes by providing a password at step 54. The user I.D. and password are used to look up applicable threshold values in step 56.

In general, threshold values are used to influence the decision of whether or not a particular digital dataset should be deemed to contain the selected category of information content. In the example at hand, threshold values are used in the determination of whether or not any particular web page should be blocked or, conversely, displayed to the user. The software can simply select a default threshold value that is thought to be reasonable for screening pornography from the average user. In a preferred embodiment, the software includes means for a parent, guardian or other administrator to set up one or more user accounts and select appropriate threshold values for each user. Typically, these will be based on the user's age, maturity, level of experience and the administrator's good judgment. The interface can be relatively simple, calling for a selection of a screening level – such as low, medium and high -- or user age groups. The software can then translate these selections into corresponding rating numbers.

Operation.

In operation, the user first logs-in with a user I.D. and password, as noted, and then interacts with the browser software in the conventional manner to "surf the web" or access any selected web site or page, for example, using a search engine or a predetermined URL. When a target page is downloaded to the user's computer, it is essentially "intercepted" by the proxy server 10, and the HTML page 12 is then analyzed as described above, to determine a rating score shown at path 46 in FIG. 5. In step 60, the software then compares the downloaded page rating to the threshold values applicable to the present user. In a preferred embodiment, the higher the rating the more likely the page contains pornographic content. In other words, a higher frequency of occurrence of "naughty" words (those with positive weights) drives the ratings score higher in a positive direction. Conversely, the presence of other terms having negative weights drives the score lower.

If the rating of the present page exceeds the applicable threshold or range of values for the current user, a control signal shown at path 62 controls a gate 64 so as to prevent the present page from being displayed at the browser display 52.

Optionally, an alternative or substitute page 66 can be displayed to the user in lieu of the downloaded web page. The alternative web page can be a single, fixed page of content stored in the software. Preferably, two or more alternative web pages are available, and an age-appropriate alternative web page is selected, based on the user I.D. and threshold values. The alternative web page can explain why the downloaded web page has been blocked, and it can provide links to direct the user to web pages having more appropriate content. The control signal 62 could also be used to take any other action based on the detection of a pornographic page, such as sending notification to the administrator. The administrator can review the page and, essentially, overrule the software by adding the URL to a "do not block" list maintained by the software.

Formulating Weighted Lists of Words and Phrases.

FIG. 6 is a simplified block diagram of a neural-network architecture for developing lists of words and weightings according to the present technology. Here, training data 70 can be any digital record or dataset, such as database records, e-mails, HTML or other web pages, use-net postings, etc. In each of these cases, the records include at least some text, *i.e.*, strings of ASCII characters, that can be identified to form regular expressions, words or phrases. We illustrate the technology by describing in greater detail its application for detecting pornographic content of web pages. This description should be sufficient for one skilled in the art to apply the principles of the technology to other types of digital information.

In FIG. 6, a simplified block diagram of a neural-network shows training data 70, such as a collection of web pages. A series of words, phrases or other regular expressions is extracted from each web page and input to a neural-network 72. Each of the terms in the list is initially assigned a weight at random, reflected in a weighted list 78. The network analyzes the content of the training data, as further explained below, using the initial weighting values. The resulting ratings are compared to the predetermined designation of each sample as "yes" or "no," *i.e.*, pornographic or not pornographic, and error data is accumulated. The error

information thus accumulated over a large set of training data, say 10,000 web pages, is then used to incrementally adjust the weightings. This process is repeated in an interactive fashion to arrive at a set of weightings that are highly predictive of the selected type of content.

FIG. 7 is a flow diagram that illustrates the process for formulating weighted lists of expressions -- also called target attribute set -- in greater detail. Referring to FIG. 7, a collection of "training pages" 82 is assembled which, again, can be any type of digital content that includes ASCII words but for illustration is identified as a web page. The "training" process for developing a weighted list of terms requires a substantial number of samples or "training pages" in the illustrated embodiment. As the number of training pages increases, the accuracy of the weighting data improves, but the processing time for the training process increases non-linerally. A reasonable tradeoff, therefore, must be selected, and the inventors have found in the presently preferred embodiment that the number of training pages (web pages) used for this purpose should be at least about 10 times the size of the word list. Since a typical web page contains on the order of 1,000 natural language words, a useful quantity of training pages is on the order of 10,000 web pages.

Five thousand web pages 84 should be selected as examples of "good" (*i.e.*, not pornographic) content and another 5,000 web pages 86 selected to exemplify "bad" (*i.e.*, pornographic) content The next step in the process is to create, for each training page, a list of unique words and phrases (regular expressions). Data reflecting the frequency of occurrence of each such expression in the training pages is statistically analyzed 90 in order to identify those expressions that are useful for discriminating the pertinent type of content. Thus, the target attribute set is a set of attributes that are indicative of a particular type of content, as well as attributes that indicate the content is NOT of the target type. These attributes are then ranked in order of frequency of appearance in the "good" pages and the "bad" pages.

The attributes are also submitted to a Correlation Engine which searches for correlations between attributes across content sets. For example, the word "breast"

appears in both content sets, but the phrases "chicken breast" and "breast cancer" appear only in the Anti-Target ("good") Content Set. Attributes that appear frequently in both sets without a mitigating correlation are discarded. The remaining attributes constitute the Target Attribute Set.

FIG. 8 illustrates a process for assigning weights to the target attribute set, based on the training data discussed above. In Figure 4, the weight database 110 essentially comprises the target attribute set of expressions, together with a weight value assigned to each expression or term. Initially, to begin the adaptive training process, these weights are random values. (Techniques are known in computer science for generating random—or at least good quality, pseudo-random—numbers.) These weighting values will be adjusted as described below, and the final values are stored in the database for inclusion in a software product implementation of the technology. Updated or different weighting databases can be provided, for example via the web.

The process for developing appropriate weightings proceeds as follows. For each training page, similar to FIG. 5, the page is scanned to identify regular expressions, and these are checked against the database 110 to form a match list 114. For the expressions that have a match in database 110, the corresponding weight is downloaded from the database and combined with the list of expressions to form a weighted list 120. This process is repeated so that weighted lists 120 are formed for all of the training pages 100 in a given set.

Next, a threshold value is selected—for example, low, medium or high value—corresponding to various levels of selectivity. For example, if a relatively low threshold value is used, the system will be more conservative and, consequently, will block more pages as having potentially pornographic content. This may be useful for young children, even though some non-pornographic pages may be excluded. Based upon the selected threshold level 122, each of the training pages 100 is designated as simply "good" or "bad" for training purposes. This information is stored in the rated lists at 124 in FIG. 8 for each of the training pages.

A neural-network 130 receives the page ratings (good or bad) via path 132 from the lists 124 and the weighted lists 120. It also accesses the weight database 110. The neural-network then executes a series of equations for analyzing the entire set of training pages (for example, 10,000 web pages) using the set of weightings (database 110) which initially are set to random values. The network processes this data and takes into account the correct answer for each page—good or bad—from the list 124 and determines an error value. This error term is then applied to adjust the list of weights, incrementally up or down, in the direction that will improve the accuracy of the rating. This is known as a feed-forward or back-propagation technique, indicated at path 134 in the drawing. This type of neural-network training arrangement is known in prior art for other applications. For example, a neural-network software packaged called "SNNS" is available on the internet for downloading from the University of Stuttgart.

Following are a few entries from an illustrative list of regular expressions along with neural-net assigned weights:

18[\W]?years[\W]?of[\W]?age[\W]	500
adults[\W]?only[\W]	500
bestiality[\W]	250
chicken[\W]breasts?[\W]	-500
sexually[\W]?(oriented|explicit)[\W]	500

Other Applications.

As mentioned above, the principles of the present technology can be applied to various applications other than web-browser client software. For example, the present technology can be implemented as a software product for personal computers to automatically detect and act upon the content of web pages as they are viewed and automatically "file," *i.e.*, create records comprising meta-content references to that web-page content in a user-modifiable, organizational and presentation schema.

Another application of the technology is implementation in a software product for automatically detecting and acting upon the content of computer files and directories. The software can be arranged to automatically create and record meta-content references to such files and directories in a user-modifiable, organizational and presentation schema. Thus, the technology can be applied to help end users quickly locate files and directories more effectively and efficiently than conventional directory-name and key-word searching.

Another application of the technology is e-mail client software for controlling pornographic and other potentially harmful or undesired content and e-mail. In this application, a computer program for personal computers is arranged to automatically detect and act upon e-mail content—for example, pornographic e-mails or unwanted commercial solicitations. The program can take actions as appropriate in response to the content, such as deleting the e-mail or responding to the sender with a request that the user's name be deleted from the mailing list.

The present technology can also be applied to e-mail client software for categorizing and organizing information for convenient retrieval. Thus, the system can be applied to automatically detect and act upon the content of e-mails as they are viewed and automatically file meta-content references to the content of such e-mails, preferably in a user-modifiable, organizational and presentation schema.

A further application of the technology for controlling pornographic or other undesired content appearing in UseNet news group postings and, like e-mail, the principles of the present technology can be applied to a software product for automatically detecting and acting upon the content of UseNet postings as they are received and automatically filing meta-content references to the UseNet postings in a user-modifiable, organizational and presentation schema.

Claims

1. A method for profiling a user of the Internet according to predefined categories of interest, comprising the steps of:

scanning content information of an Internet user to generate unknown data;

processing the unknown data to determine its relevance to predefined categories of interest; and

generating a match of the unknown data with the predefined categories to form a profile of the user.

2. A method as recited in Claim 1 wherein the profile of the user provides a level of interest in the predefined category.

3. A method as recited in Claim 1 wherein the relevance of unknown data to predefined categories is measured in a matching rating system.

4. A method as recited in Claim 3 wherein the matching rating system analyzes the unknown data according to length of time reviewing the content information.

5. A method as recited in Claim 3 wherein the matching rating system analyzes the unknown data according to frequency of encounter.

6. A method as recited in Claim 3 wherein the matching rating system analyzes the unknown data according to a statistical measure of how recently the matching information was reviewed.

7. A method as recited in Claim 3 wherein the matching rating system analyzes the unknown data according to an aging algorithm.

8. A method as recited in Claim 3 wherein the matching rate indicates strength of the unknown data with respect to the predefined categories.

9. A method as recited in Claim 3 wherein the matching rate indicates closeness of the unknown data with respect to the predefined categories.

10. A method as recited in Claim 1 wherein one of the predefined categories is related to one of sports, games, business, investing, health, hobbies, technology, arts, politics, social issues, weather and news.

11. A method as recited in Claim 1 wherein the content information is in the form of e-mail.

12. A method as recited in Claim 1 wherein the content information is in the form of Web pages.

13. A method as recited in Claim 1 wherein the content information is in the form of chat streams.

14. A method as recited in Claim 1 wherein the content information is in the form of UseNet information.

15. A method as recited in Claim 1 wherein the content information is in the form of push information.

16. A method as recited in Claim 1 wherein the method is carried out by an Internet Service Provider.

17. A method as recited in Claim 1 wherein the user has an Internet communication device and the method is carried out on the user's Internet communication device.

18. A method as recited in Claim 1 wherein the user communicates on the Internet through an Internet hub and the method is carried out at the hub.

19. A method as recited in Claim 1 further comprising reporting the profile to a tracker that operates in conjunction with an offer manager to prepare and dispense offers to the user based upon the user profile.

20. A method of forming a recognizer for use in profiling Internet user interests, comprising the steps of:

collecting representative data sets of major areas of interests; and

processing the data sets by algorithms and weighted rules to form a recognizer.

21. A method of using the recognizer formed as recited in claim 20 in furtherance of forming a profile of the Internet user, comprising the steps of:

scanning content information responsive to the use by an Internet user to form unknown data;

processing the unknown data against a recognizer; and

generating a match of the unknown data with the recognizer to form a profile of the user.

22. A method as recited in claim 20 wherein a plurality of recognizers are formed.

23. A method as recited in claim 20 wherein the recognizer is related to one of sports, games, business, investing, health, hobbies, technology, arts, politics, social issues, weather and news.

24. A method as recited in claim 21 wherein the processing step comprises:

breaking down the content information into discrete pieces of data; and

passing the discrete pieces of data into the recognizer.

25. A method as recited in claim 21 wherein the processing step occurs in real time.

26. A method as recited in claim 21 wherein the processing step is delayed.

27. A method as recited in claim 21 wherein the profile is stored on the client installation.

28. A method as recited in claim 21 wherein the profile is stored on the server installation.

29. A method as recited in claim 21 wherein the processing of unknown data is circumventable by the user.

30. A system for profiling a user of the Internet according to predefined categories of interest, comprising:

a processor configured to scan content information of an Internet user to generate unknown data;

a processor configured to process the unknown data to determine its relevance to predefined categories of interest; and

a processor configured to generate a match of the unknown data with the predefined categories to form a profile of the user.

31. A system for forming a recognizer for use in profiling Internet user interests, comprising:

a processor configured to collect representative data sets of major areas of interests to form data sets; and

a processor configured to process the data sets by algorithms and weighted rules to form a recognizer.

32. A system as recited in claim 31 using the recognizer in furtherance of forming a profile of the Internet user, comprising:

a processor configured to scan content information responsive to the use of such content by an Internet user to generate unknown data;

a processor configured to process the unknown data against a recognizer; and

a processor configured to generate a match of the unknown data with the recognizer to form a profile of the user.

33. A system for profiling a user of the Internet according to predefined categories of interest, comprising:

a processor configured to scan content information of an Internet user to generate unknown data;

a processor configured to process the unknown data to determine its relevance to predefined categories of interest;

a processor configured to generate a match of the unknown data with the predefined categories to form a profile of the user; and

a processor configured to report the profile to a tracker that operates in conjunction with an offer manager to prepare and dispense offers to the user based upon the user profile.

34. A system as recited in Claim 33, wherein the offers dispensed to the user are delivered via the Internet.

FIG. 1

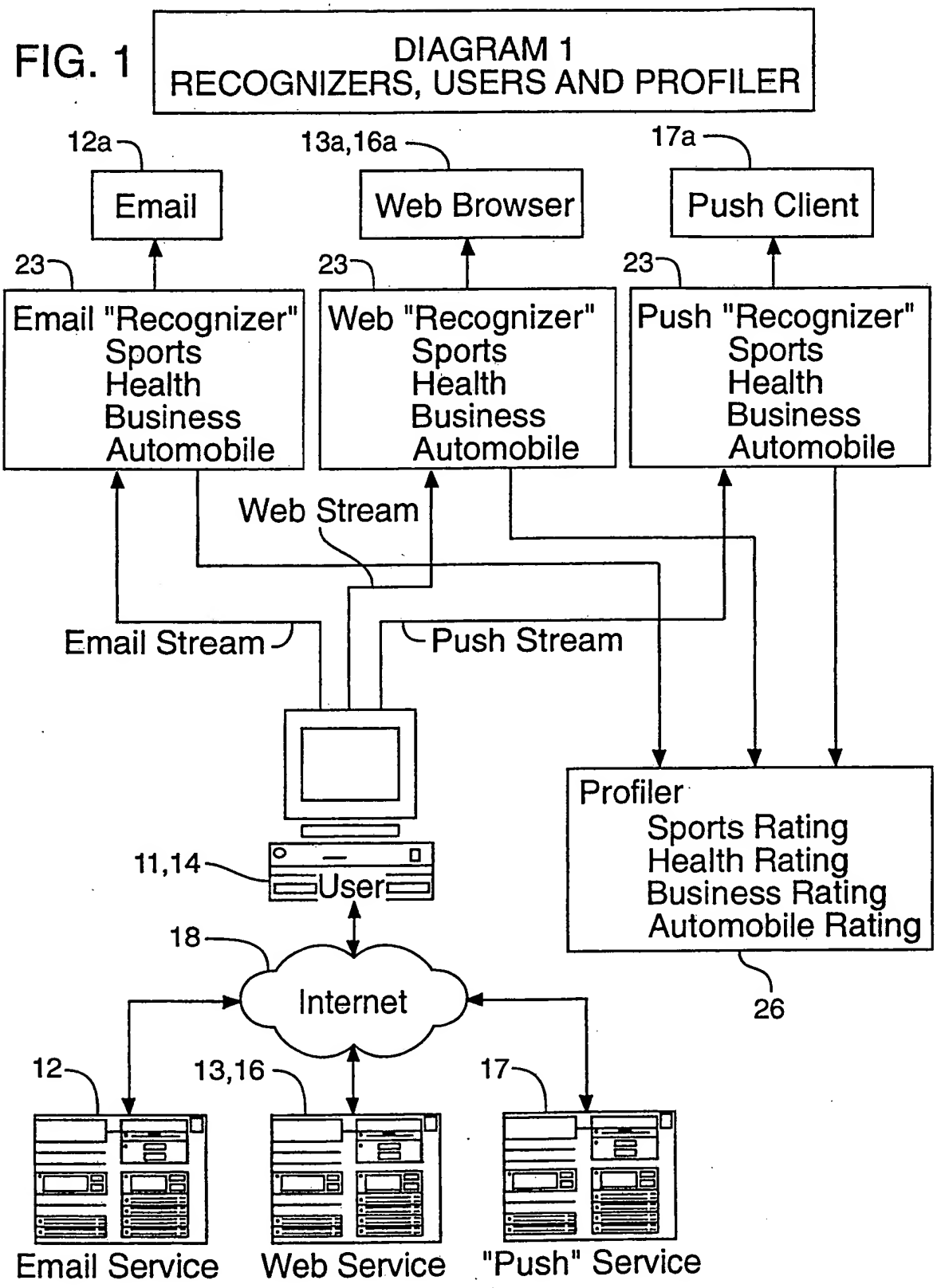


DIAGRAM 1
RECOGNIZERS, USERS AND PROFILER
12a
Email
13a,16a
Web Browser
17a
Push Client
23
Email "Recognizer"
Sports
Health
Business
Automobile
23
Web "Recognizer"
Sports
Health
Business
Automobile
23
Push "Recognizer"
Sports
Health
Business
Automobile
Web Stream
Email Stream
Push Stream
11,14
User
Profiler
Sports Rating
Health Rating
Business Rating
Automobile Rating
26
18
Internet
12
13,16
17
Email Service
Web Service
"Push" Service

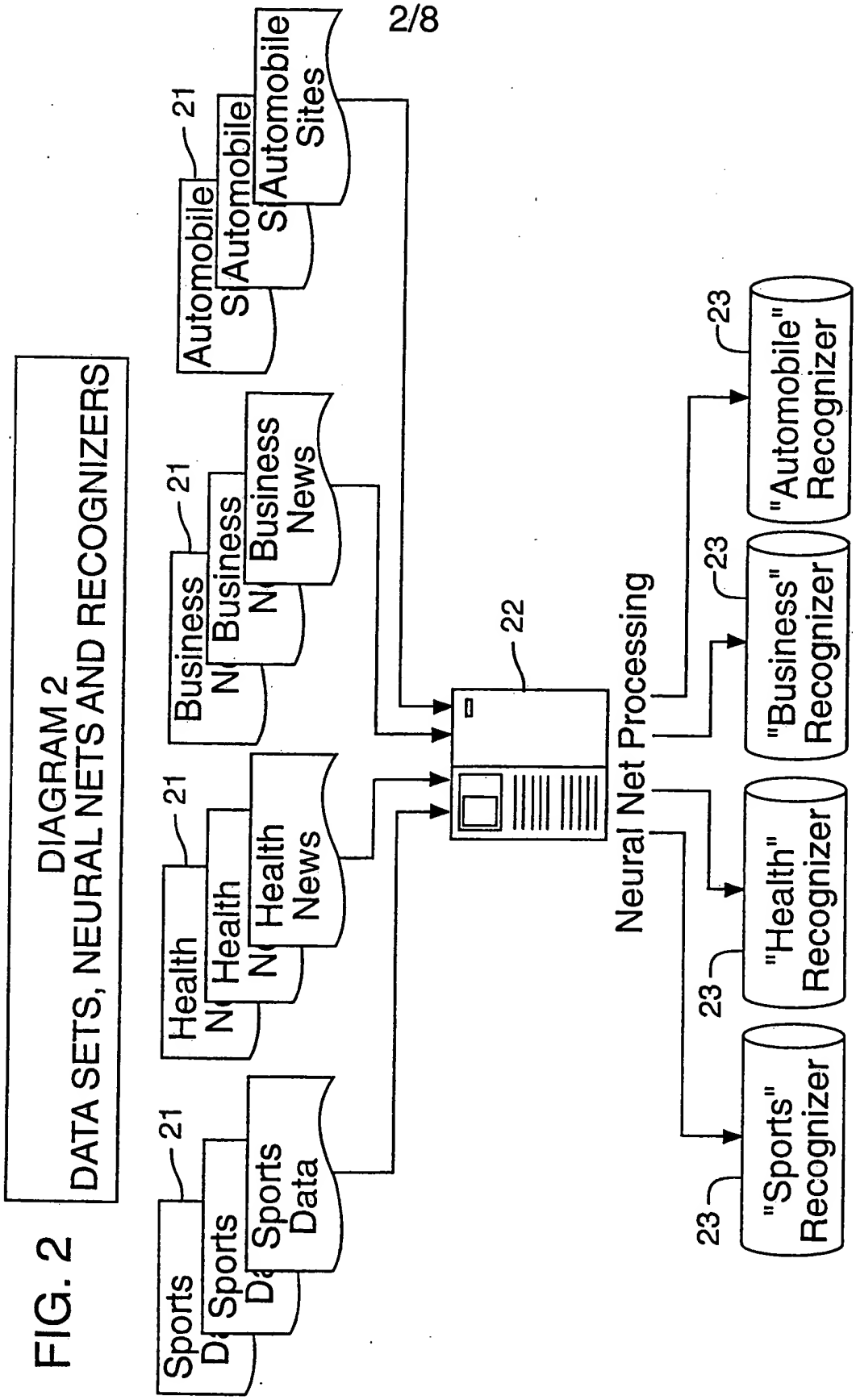


DIAGRAM 2
DATA SETS, NEURAL NETS AND RECOGNIZERS
Sports Data
Health News
Business News
Automobile Sites
21
22
Neural Net Processing
23
"Sports" Recognizer
"Health" Recognizer
"Business" Recognizer
"Automobile" Recognizer

FIG. 2

FIG. 3

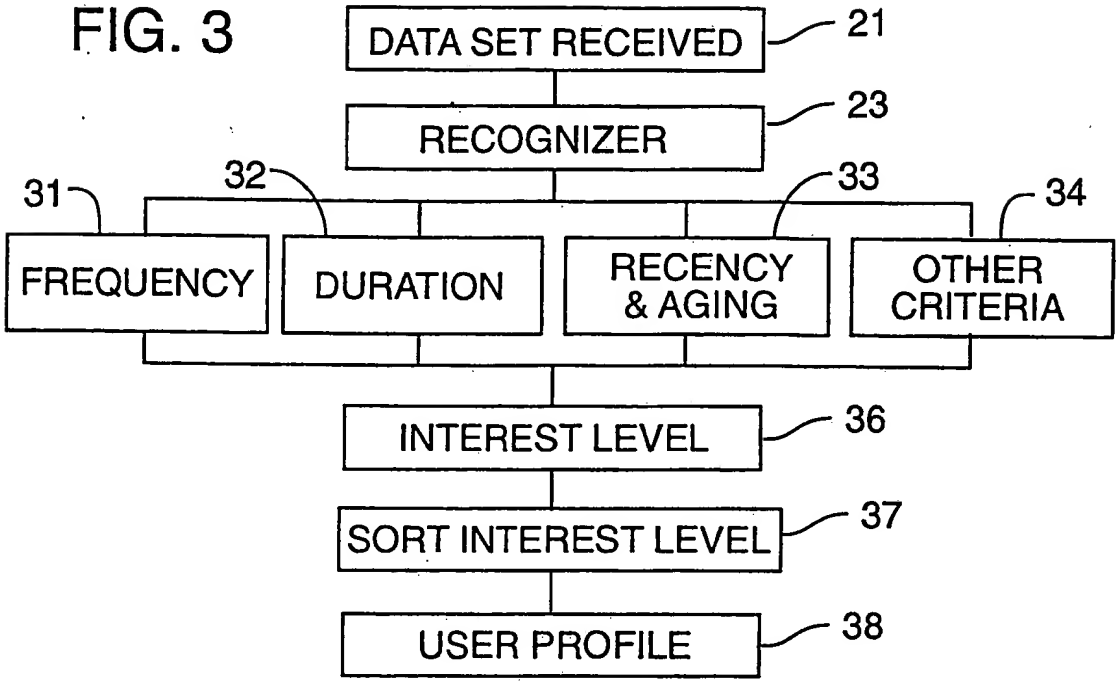

DATA SET RECEIVED
21
RECOGNIZER
23
31
32
33
34
FREQUENCY
DURATION
RECENCY & AGING
OTHER CRITERIA
INTEREST LEVEL
36
SORT INTEREST LEVEL
37
USER PROFILE
38

FIG. 4

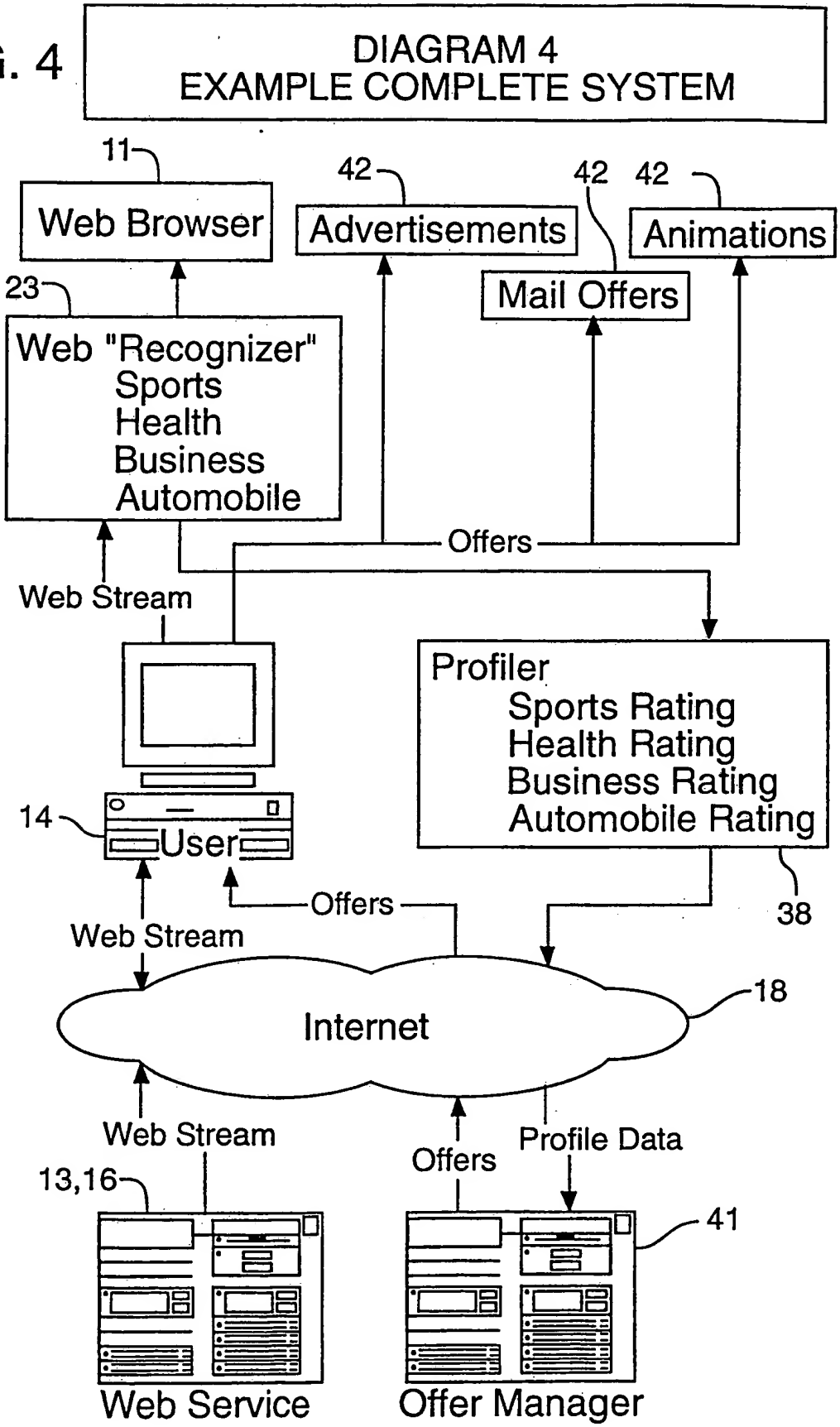


DIAGRAM 4
EXAMPLE COMPLETE SYSTEM
11
Web Browser
42
Advertisements
42
Mail Offers
42
Animations
23
Web "Recognizer"
Sports
Health
Business
Automobile
Offers
Web Stream
Profiler
Sports Rating
Health Rating
Business Rating
Automobile Rating
14
User
Offers
Web Stream
38
Internet
18
Web Stream
Offers
Profile Data
13,16
41
Web Service
Offer Manager

FIG. 5

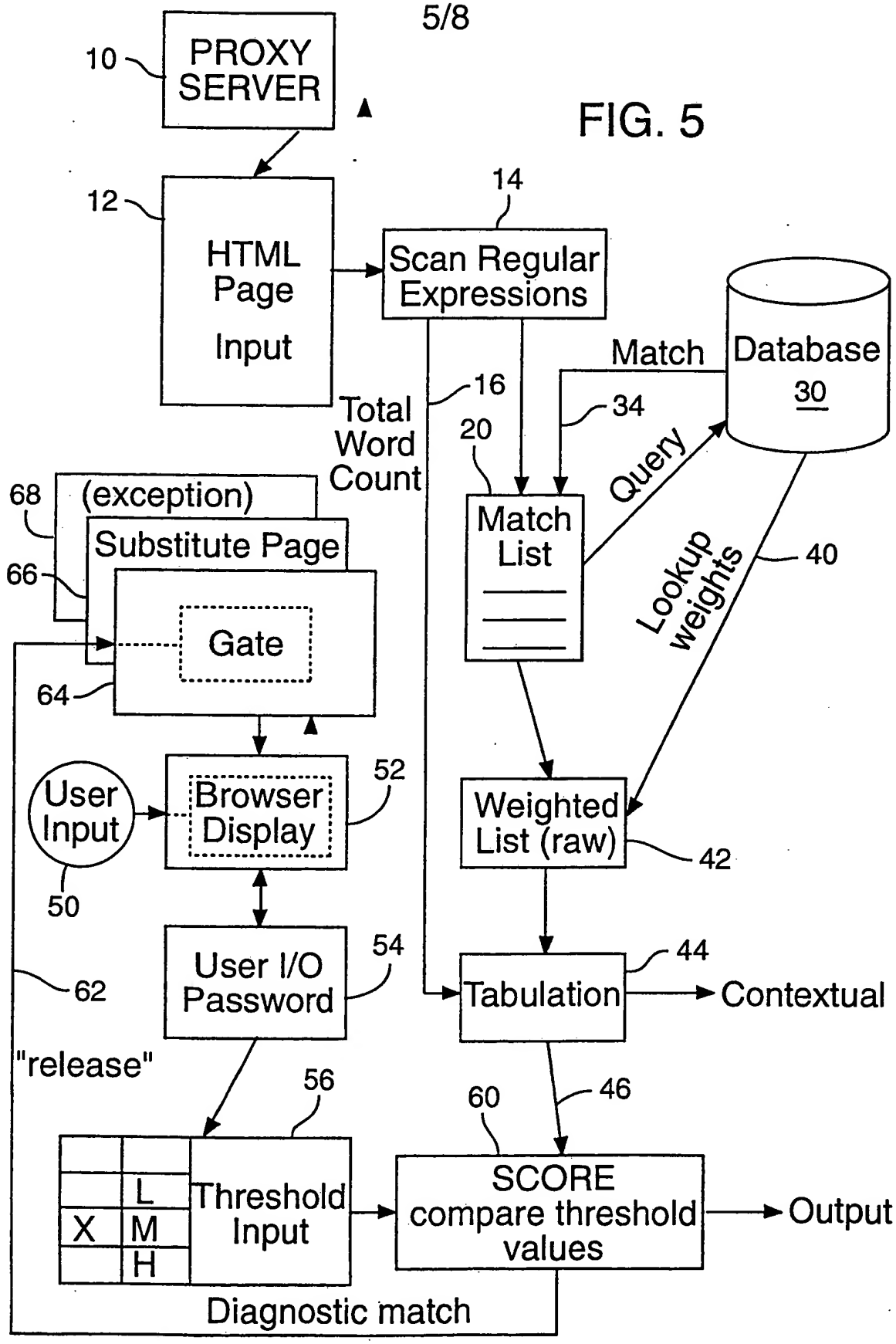

10
PROXY SERVER
12
HTML Page Input
14
Scan Regular Expressions
Database 30
Match
16
Total Word Count
34
20
Query
68
(exception)
Substitute Page
66
Gate
64
Match List
Lookup weights
40
User Input
50
Browser Display
52
Weighted List (raw)
42
User I/O Password
54
62
Tabulation
44
Contextual
"release"
56
60
46
X
L
M
H
Threshold Input
SCORE compare threshold values
Output
Diagnostic match

FIG. 6

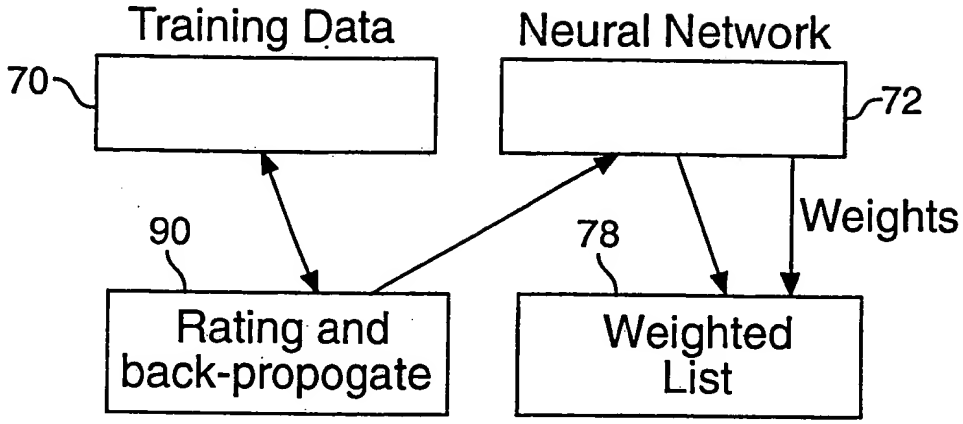
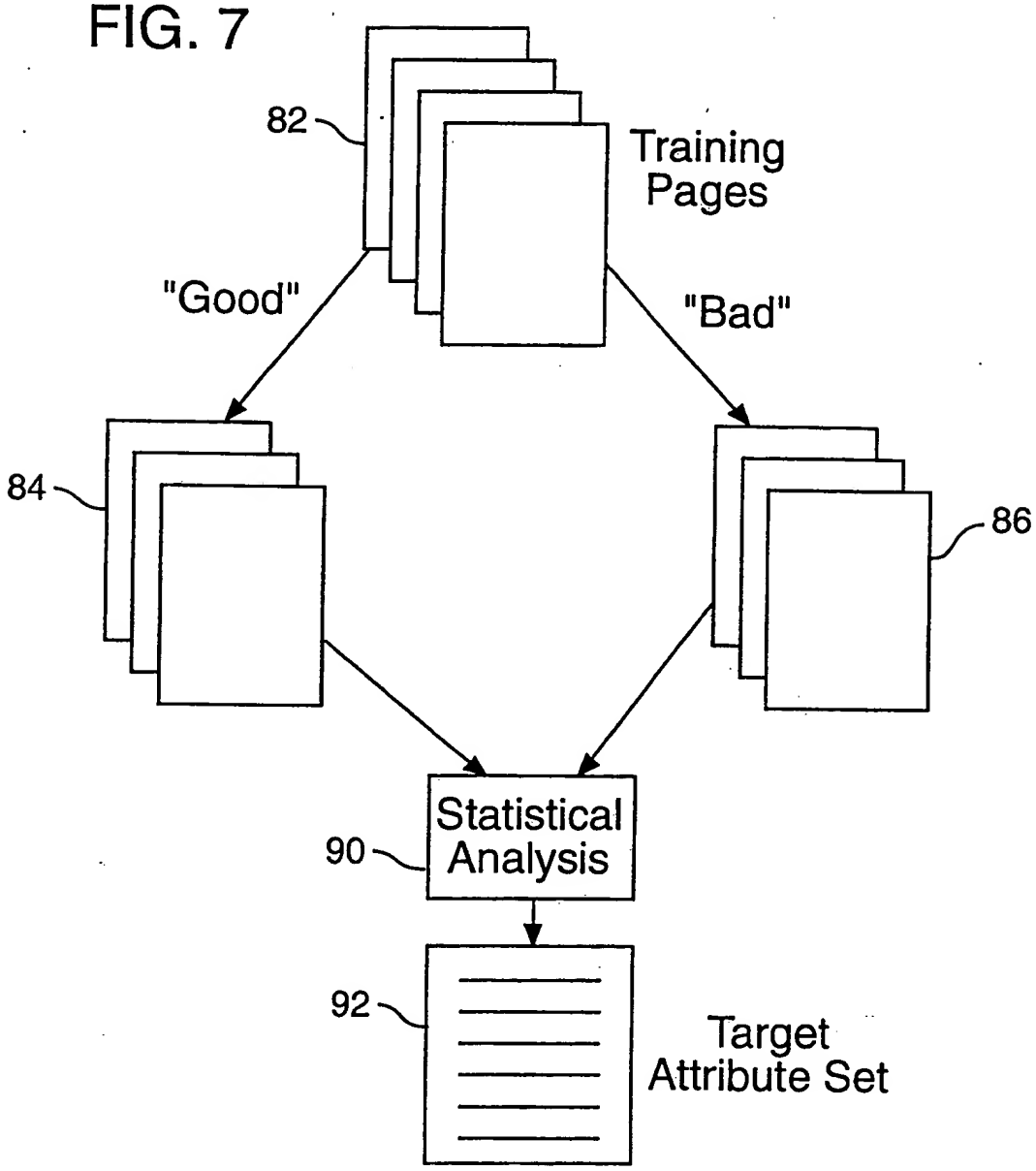
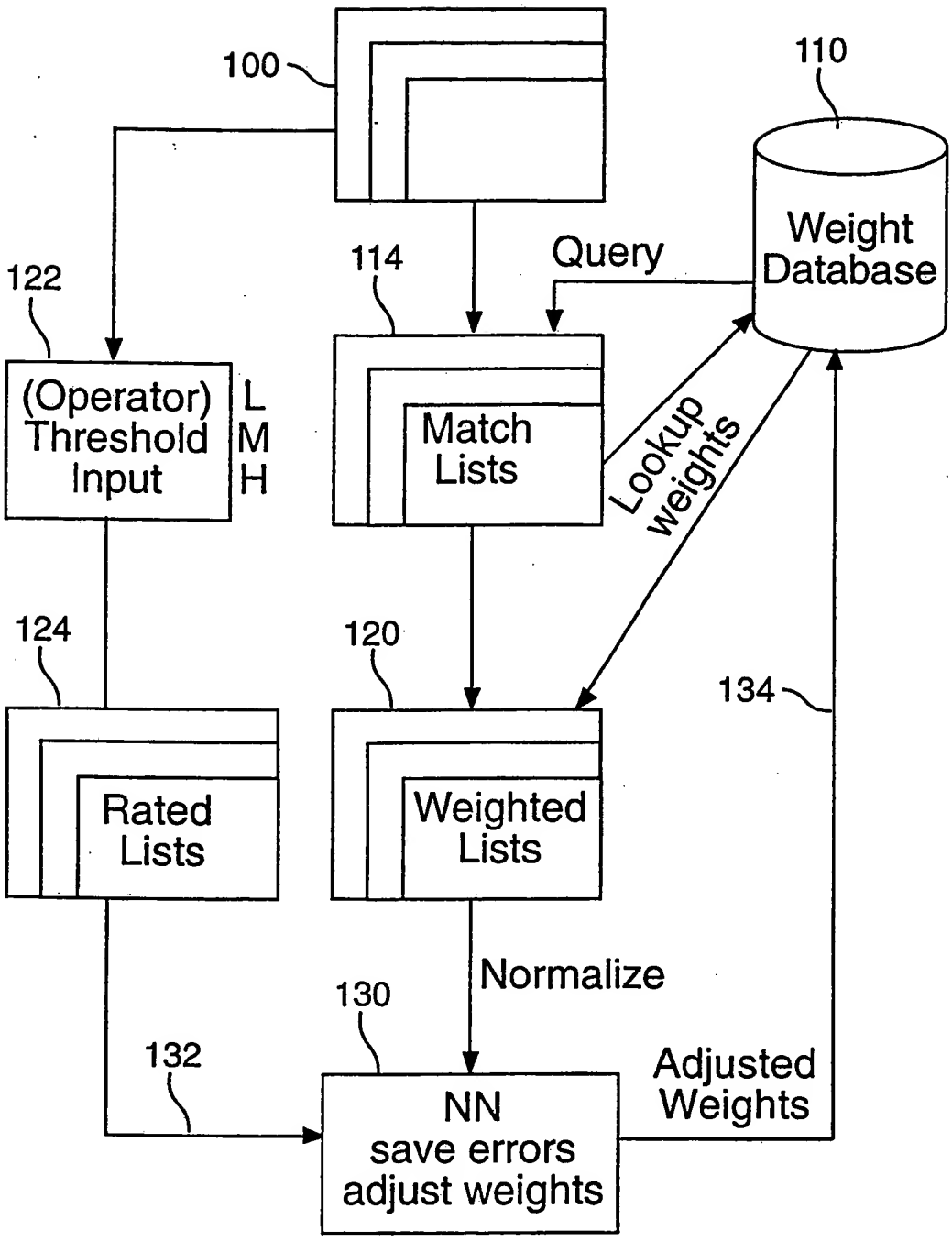

Training Data
Neural Network
70
72
Weights
90
78
Rating and back-propogate
Weighted List

FIG. 7


82
Training Pages
"Good"
"Bad"
84
86
Statistical Analysis
90
92
Target Attribute Set


100
110
Weight Database
Query
114
122
(Operator) Threshold Input
L
M
H
Match Lists
Lookup weights
124
120
134
Rated Lists
Weighted Lists
Normalize
130
132
Adjusted Weights
NN
save errors
adjust weights

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US99/17654

A. CLASSIFICATION OF SUBJECT MATTER
IPC(6) :GO 6F 17/30
US CL : 707/5
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 707/5; 707/1-4,6,10; 709/217-219

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

APS, IEEE

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 5,446,891 A (KAPLAN et al) 29 August 1995, Abstract	1-34
Y	US 5,537,586 A (AMRAM et al) 16 July 1996, Fig. 3 and 4	1-34
Y	US 5,727,129 A (BARRETT et al.) 10 March 1998, Abstract, Fig. 4-7	1-34
Y	US 5,754,938 A (HERZ et al.) 19 May 1998, Abstract, columns 55-63	1-34
X, P	US 5,790,935 A (PAYTON) 04 August 1998, column 9, lines 42-48; Fig. 3b, 3c, 4, and 7b	1-9, 20-22, 24, 30-34
Y, P		10-19, 23, 25-29

[X] Further documents are listed in the continuation of Box C. [ ] See patent family annex.

*	Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be of particular relevance	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"B"	earlier document published on or after the international filing date	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&"	document member of the same patent family
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search	Date of mailing of the international search report
29 SEPTEMBER 1999	18 OCT 1999
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703) 305-3230	Authorized officer HOSAIN T. ALAMJames R. Matthews Telephone No. (703) 308-6662

Form PCT/ISA/210 (second sheet)(July 1992)*

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US99/17654

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	KHAN et al., Categorizing Web Documents Using Competitive Learning: An Ingredient of a Personal Adaptive Agent, International Conference on Neural Networks, June 1997, Vol. 1, pages 96-99	1-2, 4-9, 20-22, 24-25, 27-34
Y		3, 10-19, 23, 26

Form PCT/ISA/210 (continuation of second sheet)(July 1992)*